STATISTICA, anno LXVIII, n. 2, 2008

ON SOME COUNTER-COUNTER-EXAMPLES ABOUT CLASSICAL INFERENCE

Benito V. Frosini

1. INTRODUCTION. STATE OF CONTENTION

All statisticians, and many practitioners, have some knowledge of the existence, and contraposition, of several approaches – and corresponding techniques – in making inference from random samples to populations or random variables; and also that such contraposition is especially concerned with the two main streams of statistical inference: (a) the "sampling-based inference", following the works and legacy of Fisher and Neyman-Pearson, and (b) the Bayesian inference, founded on a preliminary assessment of a probability distribution over the possible hypotheses. A large subset of the above set of statisticians and practitioners is also aware of the strange asymmetry existing between the direction of criticisms coming from one field toward the other: there are several papers and books, written by Bayesian scholars, pointing to the presumed inadequacy of the opposite field (some major contributions will be commented on in the sequel), while the other kind of criticism is almost a curiosity (see Efron, 1986). Actually, most Non-Bayesians simply neglect, or simply do not worry, about the methodological criticisms arising from the Bayesian field; I feel very sympathetic with this viewpoint – and perhaps it is the wisest one – but I also think that sometimes it can be useful – for the sake of clarification – to cope with these criticisms and display their fundamental weakness.

The answer to Efron's question (Efron, 1986) "Why Isn't Everyone a Bayesian?", is even too simple: many people wish *to treat a constant like a constant*, not as a random variable. If I know that a box contains tokens of two kinds A and B, and I have no idea of the (random?) process leading to filling up the box with the tokens, the proportion θ of A tokens is a constant, relative just to the given box. Now, if such information is of my personal use, nobody can prevent me from contriving a more ore less (for me) reliable evaluation of the possible values of θ. However, if my inference (usually based on a random sample of the tokens drawn from the box, if a complete census is excluded) is directed towards other people, who – in case of scientific communication – is the world scientific community, the following alternative is called for: (a) either we treat the proportion θ like a

constant, or (b) we propose a prior distribution *demonstrably* acceptable for the given purpose. In the first case, I see no other possibility than making use of the (sole) information provided by the random sample drawn from the population of tokens. In the second case, I should choose a prior distribution (often of improper kind, just a means to an end), such that it has an acknowledged property of being minimally affecting the posterior distribution, given the data and the sampling process. This last approach has gained ever more approval and useful results, since its starting with a famous paper by Bernardo (1979), however in the stream of forerunner contributions by Jeffreys (1939/1948), Jaynes (1968, 1976), and many followers, among which Berger has made the most significant advancements (for important survey papers see Bayarri and Berger, 2004; Bernardo, 2005; Berger, 2006). Moreover, it must be recognized that the Bayesian approach is usually easier and more direct to implement than the classical approach, thus it could be adopted by any orthodox statistician, just to achieve a valid operative proposal.

Now, if the state of affairs is the one just sketched, what is the use of discussing old counter-examples (contrived by Bayesians just to show the unsuitableness or even the incoherency of classical proposals)? I think it is time to sweep away all the junk that some Bayesians have devised to discredit classical inference, while looking for a peaceful living together.

A warning is perhaps expedient before entering the debate. It is well known that some frequentist procedures, commented on in the sequel, could be dealt with by conditioning on a suitable ancillary statistic, thus leading to a new procedure – always in the frequentist domain – exactly or practically overlapping a standard Bayesian procedure. Three classical examples are the following: (a) conditioning on the experiment actually performed, chosen at random in the first stage of a two-stage experiment (Cox, 1958, p. 360; Cox-Hinkley, 1974, pp. 32, 38; Berger-Wolpert, 1988, p. 6; Frosini, 1991, p. 559); (b) conditioning on the marginal frequencies of a 2×2 contingency table, leading to Fisher's *exact test* (Fisher, 1935; Lindgren, 1962, p. 338); (c) conditioning on the *configuration* of a sample (vector of differences between successive order statistics), when the inference concerns the location parameter θ in a location family with density $f(x - \theta)$ (Fisher, 1934; Cox-Hinkley, 1974, pp. 34, 221).

While conditioning to ancillary in case (a) would probably be followed by most people, my feeling is that a decreasing proportion is likely to be encountered when passing to (b) and (c) cases. I think I can confirm an opinion already expressed: "... the choice of the *reference set* within a frequentist decisional approach cannot be left to the researcher imagination; on the contrary, such a choice must be bound with the decision problem concretely defined, and with the actual or theoretical possibility of repeating the random experiment – under homogeneity conditions – for a sufficiently large or simply unbounded number of times (Frosini, 1999, pp. 166-7; see also Cox-Hinkley, 1974, p. 116).

Having said that, it must be stressed that the counter-examples in the sequel will be just *countered* on the basis of *logical* arguments, with the aim of showing that they are logically unacceptable. Pointing to conditional procedures as alternative

frequentist solutions could obscure the essential fact that the counter-examples themselves are untenable. Other non-Bayesian approaches (e.g. conditioning, empirical Bayes) could be examined all the same for possible alternative solutions, but they are beyond the scope of this paper.

2. OBJECTIVE BAYESIAN INTERVALS AND FAILURE OF BAYESIAN INFERENCE

The topic commented on in this Section is the only one of the paper which is not a counter-counter-example, but only a counter-example: it is so fundamental and plainly true that it is hardly touched on explicitly even by opponents of Bayesianism; its comprehension is best ensured in the *ideal case* of Bayesian reasoning, i.e. when an effective random distribution for the parameter θ exists (the case of a constant θ – with probability one – is included as a limiting case). The "objectivity" thus assumed lies in the *existence* – and *knowledge* – of a two-stage experiment: the first stage yields a *random* value for the parameter θ, depending on a density $f(\theta)$, $\theta \in \Theta$; the second stage yields a *random* value $x \in \Omega$ of a variable X, depending on a density (likelihood) $g(x \mid \theta)$ (for the sake of simplicity, we maintain the term "density" for any kind of random variable) (Frosini, 2005, pp. 437-438).

When some wrong distribution is introduced over the set of hypotheses, it can *usually* lead – but *not necessarily* – to more or less gross mistakes in the inferential process. The most unfortunate case is perhaps when θ is not random at all; thus the "random" variable θ is actually degenerate in a constant $\theta = \theta_0$, taken with probability one. In such a case all *objective* evaluations of an interval I – hopefully comprising θ_0 – are of the following kind (being T any statistic):

$$\begin{aligned} P(\theta \in I \mid T = t) &= 1 \quad \text{if } \theta_0 \in I \\ &= 0 \quad \text{if } \theta_0 \notin I \end{aligned}$$

as the degenerate θ is independent of any random variable T. Of course, if θ is the object of inference, θ_0 is unknown, and we can confidently obtain an interval I endowed with coverage probability one only if we equate I with (a superset of) Θ, the set of all possible θ values (assuming that Θ is chosen sufficiently large as to include θ_0).

With this example we have begun to answer a question of the kind raised by Jaynes (1976, p. 207) and many other scholars of the Bayesian School: "To the best of my knowledge, nobody has ever produced an example where the Bayesian method fails to yield a reasonable result", however replacing *reasonable* with *correct*, or *right*. An explanation is called for.

It is quite obvious that, putting together all the (estimated) relevant information for a given inferential problem, and applying generally approved rules of deduction and induction, I can be sufficiently satisfied: I have made the best of the available information and of my personal expertise. However, excepting perhaps a solipsist Bayesian, who asks no more than this, practically all people consider

such an elaboration only as a means to an end: what really matters is to achieve a correct, or right, result. The mere *coherency* of the implemented procedure is unable to ensure a correct result. Coherency, in itself, is worthless. This is the reason of the above replacement: instead of "reasonable result", which can be referred only to the elaboration stage, it matters to speak of "correct result", which can be recognized only by a comparison of the performed inference and the state of the world (to which the inference is directed – see Cox, 1986, p. 125; Frosini, 1989, p. 225). In this sense I wholeheartedly subscribe Jaynes statement: "*The merits of any statistical method are determined by the results it gives when applied to specific problems*" (italics in the original) (Jaynes, 1976, p. 178).

With this clarification, we can briefly quote two other kinds of controls reported by Frosini (2005). The first kind of control considers the classical case of normal prior for the mean λ of a normal distribution, and normal likelihood; in the quoted paper the *contrived* normal prior $N(\lambda_1,\sigma_1^2)$ for λ is flanked by the *real* normal prior $N(\lambda_0,\sigma_0^2)$. By giving $\lambda_1 \neq \lambda_0$ and/or $\sigma_1 \neq \sigma_0$, the coverage probability (computed by assuming the *wrong* prior) of the correct 95% Bayesian interval (computed on the *true* prior distribution) can easily yield values much smaller than the nominal 95%; thus, the *result* obtained by applying a Bayesian procedure can be totally wrong. A similar inquiry, however aimed at showing an increase of the expected squared error risk of the Bayes estimator of λ, was made by Efron and Morris (1971). Another kind of application has been examined by Frosini (2005), where the likelihood remains normal, however with a prior for the mean which is uniform over a finite interval. Also in this case, maintaining the same kind of prior distribution but changing the interval, the coverage probabilities (assuming the wrong prior) of the correct 95% Bayesian interval (derived from the true prior distribution) have been computed, and some *appalling results* have been observed. *In all the cases*, of course, the classical confidence intervals do their duty: the *true* unknown parameter is always included in the confidence interval 95% of the time.

3. FLAT LIKELIHOOD AND UNBIASED LINEAR ESTIMATORS IN SAMPLING FROM FINITE POPULATIONS

A curious result by Godambe (1955, 1965), which has been generally misapprehended in the literature, regards the so-called *flat likelihood* in survey sampling; it seems to support the Bayesians' claim that the sampling plans are irrelevant (in order to make inferences from the sample). This result is possible if we start correctly with the definition of a sampling plan p as the collection of all the admissible samples s – containing one or more objects of the population – together with their probabilities $P(s)$. If the N objects or units in the population are identified by the *labels* 1, 2, ..., N, and the study variable X (usually multidimensional) takes the corresponding values $X_1, X_2, \ldots, X_N$, the vector $\boldsymbol{X} = (X_1, \ldots, X_N)$ is *defined* as the *parameter* of the population. Letting x_i be the *observed* value of X_i, and calling $\boldsymbol{x}_s = \{x_i : i \in s\}$ the set of variate values pertaining to the units drawn from the

population, by *sample data* we intend the couple $(s,\boldsymbol{x}_s) \equiv \{(i,x_i) : i \in s\}$. By resuming the exposition in Frosini (1996*a*, pp. 218-221), being $P\{(s,\boldsymbol{x}_s)\} = P(s)$, the *likelihood* of the sample s, as a function of the parameter $\boldsymbol{X}$, is *defined* as

$$L\{\boldsymbol{X} \mid (s,\boldsymbol{x}_s)\} \propto P(s) \text{ for } \boldsymbol{X} \in R^N(x_i : i \in s) \\ = 0 \text{ otherwise} \tag{1}$$

where $R^N(x_i : i \in s)$ is the subset of the euclidean space R^N such that, for the coordinates X_i, $i \in s$, $X_i = x_i$ holds (in other words, it is the set of parameter points $\boldsymbol{X}$ that are consistent with the sample data $\boldsymbol{x}_s$) (Godambe, 1969, p. 249). Therefore, such a *likelihood* is flat on the entire parametric space $R^N(x_i : i \in s)$ consistent with the sample.

This impressive result essentially depends on defining by *sample data* the couple $(s,\boldsymbol{x}_s)$, namely by taking into account not only the data $\boldsymbol{x}_s$, but also the n individuals sampled. Writes Godambe (1965): "The characteristic difference of the populations, we come across in sample surveys, from other populations, is this: *Here apart from the variate values, units having those variate values are identifiable.* This fact has mostly been overlooked.". True. However, this possibility of identification, or labelling, has been given a meaning wholly unjustified, because the individuals, inside the population or the stratum from which they are drawn, are "exchangeable"

It is well known that the concept of likelihood considered in traditional inference is at variance with the one just sketched: given a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ of variate values measured on n objects, having the density $g(\boldsymbol{x},\boldsymbol{\theta})$ indexed by a parameter $\boldsymbol{\theta}$ belonging to a parameter space $\boldsymbol{\Theta}$, the likelihood of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ is defined by

$$L(\boldsymbol{\theta}\,;\,\boldsymbol{x}) \propto g(\boldsymbol{x},\boldsymbol{\theta}) \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{2}$$

and – except very special cases – is not constant over $\boldsymbol{\Theta}$. There is no need of adopting different approaches for infinite and finite populations: in both cases, we are interested in the frequency distribution of a variable X, and we can attack the inference by means of the likelihood of the sample data $\boldsymbol{x}$. This clear position was taken by Royall (1968), in his comment on Godambe's proposal. It is true that we are practically never in a position to identify the possible hypotheses for the distribution of X in a finite population; thus, the inference for a finite population has an elective place in the domain of non-parametric inference. Anyway, if a list of such hypotheses were available, a likelihood-based inference is possible, as illustrated by Frosini (1996*a*, pp. 220-223).

A famous related result by Godambe (1955), usually quoted as a weakness of the traditional theory of inference for finite populations, is that an unbiased linear estimator with least variance does not exist. Let U (with elements u) be the set of units in the study population, and S (with elements s) be the set of all possible samples for a given sampling design. Godambe (1955) considers the most general linear estimator of the total $T = \sum_1^N X_i$ given by

$$e(s,\boldsymbol{X}) = \sum_{u \in s} \beta(s,u)x(u) \tag{3}$$

and shows that *in this class of estimators* it is not possible to find an estimator with minimum variance uniformly in $\boldsymbol{X}$. It is easily found (Frosini, 1996a, pp. 217, 226-229) – as Godambe himself surmised – that the result is wholly evident at first sight; on the other hand, the same result is recognized void of interest from the viewpoint of statistical inference, because it plainly states that, by providing complete information over the N couples (i,x_i), it is possible to construct an exact estimator of the total $\sum x_i$.

4. PROBABILITY INTERVALS VS CONFIDENCE INTERVALS

The most disturbing feature in reading several criticisms of Bayesians towards non-Bayesian approaches to inference is that Non-Bayesians are assumed incapable of applying the most elementary rules of logic, and even good sense. My reference is especially concerned with the odd, or foolish behaviour imputed to somebody, who deliberately discard a piece of information which is really important and influential for a decision to be made. On the contrary, we think that Carnap's *requirement of total evidence* must be accepted, and implemented in every case: "in the application of inductive logic to a given knowledge situation, the total evidence available must be taken as basis for determining the degree of confirmation" (Carnap, 1962, p. 211). And on this point I want to reach the limits of obviousness: among the certain facts known to the subject there are also the *rules of deduction*.

A formal adhesion is required: if some procedure, aimed at gaining insight for something unknown, deliberately excludes known and influential facts, *is not an inferential procedure*; it can be, at most, an "exercise". As we shall see, these exercises have usually exploited something resembling a confidence interval; it is thus necessary to make a distinction (and this will be made – for simplicity – with respect to a unidimensional parameter θ and a two-sided interval).

Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a random sample from a random variable X whose distribution depends on a parameter θ belonging to a parameter space Θ, and $I(\boldsymbol{X}) = [a(\boldsymbol{X}),b(\boldsymbol{X})]$, with $a(\boldsymbol{X}) \leq b(\boldsymbol{X})$, be a random interval, such that (for simplicity we fix an exact probability equality):

$$P_\theta\{\theta \in I(\boldsymbol{X})\} = P_\theta\{ a(\boldsymbol{X}) \leq \theta \leq b(\boldsymbol{X})\} = p \text{ for every } \theta \in \Theta. \tag{4}$$

Given the condition (4), the interval $[a(\boldsymbol{X}),b(\boldsymbol{X})]$ can be trivially defined a *probability interval* for θ with coverage probability p. No problem until this point. Now we could inquire whether the interval $[a(\boldsymbol{X}),b(\boldsymbol{X})]$, with the associated probability p, can be of some value in inferring the subset of Θ which is most likely to include θ (the specific parameter value on which the production of the sample $\boldsymbol{X}$ depends in the particular instance). The answer is made to depend on the relevant facts,

and the deductive relations, alluded to above, which are known *before* performing the random experiment (and we limit ourselves, to be sure, on *certain* facts and relations). In other words, a control is required whether a *deductive implication* exists from the logical conjunction of the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ and the parameter space Θ

$$(\boldsymbol{x}, \Theta) \Rightarrow \Theta^S \tag{5}$$

that reduces *with certainty* the original parameter space Θ to a subset Θ^S. When such is the case, there remain two major possibilities: (a) the reduction (5) is just to a singleton $\{\theta_0\}$, thus we have *deductively* made our inference beyond all hope, and the probability statement (4) is of no inferential value; (b) in general, the intersection

$$[a(\boldsymbol{x}),b(\boldsymbol{x})] \cap \Theta^S \tag{6}$$

must be performed – one particular case being simply $[a(\boldsymbol{x}),b(\boldsymbol{x})] \cap \Theta$ - and the probability p is transferred to this intersection.

That (6) is actually a confidence interval, or – more generally – that $I(\boldsymbol{X}) \cap \Theta^S$ is a confidence set, derives from the trivial equality

$$P_\theta\{(\theta \in I(\boldsymbol{X})) \cap (\theta \in \Theta^S)\} = P_\theta\{\theta \in I(\boldsymbol{X})\} = p \tag{6bis}$$

being $(\theta \in \Theta^S)$ a certain event, and as such independent on any random event; an example of this kind will be commented on concerning formula (10).

Although not really necessary, the special case of intervals could be formalized as follows. From (4), let

(A): $a(\boldsymbol{X}) \leq \theta \leq b(\boldsymbol{X})$ a probability interval with probability p,
(B): $c(\boldsymbol{X}) \leq \theta \leq d(\boldsymbol{X})$ an interval including θ with probability one (a sub-case is c and d constant, for example from a preliminary information on physical constraints),
(C): $\max[a(\boldsymbol{X}), c(\boldsymbol{X})] \leq \theta \leq \min[b(\boldsymbol{X}), d(\boldsymbol{X})]$.
Possible cases (dropping the reference to the sample $\boldsymbol{X}$ for simplicity):
(I) (A): $a \leq \theta \leq b$ is true; as a consequence, also (C) is true, for any relation between a and c, b and d;
(II) (A): $a \leq \theta \leq b$ is false. In order to get an interval including θ, one has to diminish a or to increase b. If $\theta < a$, from (B) it must be $c < a$, hence $\max(a, c) = a$, and for any relation between b and d *also* the interval (C) does not include θ. If $\theta > b$, from (B) it must be $d > b$, hence $\min(b, d) = b$, and for any relation between a and c *also* the interval (C) does not include θ.

For the easiness of disposing of an appropriate denomination, the intersection (6) can be called *confidence set*, or, more traditionally, *confidence interval* when it results in an interval (possibly degenerate, thus including case (a)). Of course, when there is no implication that is able to reduce Θ, (4) already defines the confidence interval.

As we shall see in Section 6, many counter-examples devised by Bayesians exploit a support (of the random variable under study) that depends on the unknown parameter θ (for example, $\theta \geq \theta_0$). While the Bayesian approach takes account of this fact just in assessing the prior distribution of θ, a probability interval usually does not take into account the structure or the restrictions for the parameter space Θ; thus in these cases, if we want to make use of a probability interval as a confidence interval, such kind of constraints must be expressly introduced.

The above procedure cannot exclude that the intersection considered in formula (6) or (6 bis) is empty. Example 4 in Section 6 provides a case of this kind, where the parameter θ is the mean μ of a normal distribution, with constraints $\mu_0 \leq \mu \leq \mu_1$, and the probability interval is based on the probability distribution of the sample mean. Unlikely as they can be, probability intervals $I(\boldsymbol{X})$ completely below or beyond the interval $[\mu_0 , \mu_1]$ are nonetheless possible; in these cases the intersection $[\mu_0 , \mu_1] \cap I(\boldsymbol{X})$ is empty. If one wishes – all the same – that a confidence interval (or set) must not be empty, a suitable definition could be adopted, leading – for example – to confidence intervals comprising only μ_0 or only μ_1 (cf Cox-Hinkley, 1974, pp. 224-228); this choice, however, implies an overcoverage ($> p$) of such confidence intervals when $\mu = \mu_0$ or $\mu = \mu_1$. Of course, especially when p is rather large, an empty set for the intersection (6) could suggest reasonable doubts about the correctness of the model (in particular, concerning the restriction on the parameter space). On this point, the interested reader is referred to Mandelkern (2002) (with an interesting discussion by G. Casella, L.J. Gleser, L. Wasserman, D.A. van Dyk, M. Woodroofe and T. Zhang).

5. ON THE INDUCTIVE MEANING OF REALIZED CONFIDENCE INTERVALS

Before going on, it is expedient to sweep the field of an old diatribe, that should disappear from the debate concerning inferential procedures. Let us assume for simplicity that the interval $I(\boldsymbol{X}) = [a(\boldsymbol{X}),b(\boldsymbol{X})]$ in formula (4) is actually a confidence interval with coverage probability (or coefficient) p; a sample $\boldsymbol{x}$ is obtained, and the *realized* confidence interval $I(\boldsymbol{x}) = [a(\boldsymbol{x}),b(\boldsymbol{x})]$ is computed. What sense can we attach to the statement that

$$a(\boldsymbol{x}) \leq \theta \leq b(\boldsymbol{x}) \text{ with confidence } p? \tag{7}$$

Of course, an *objective* probability for this statement (θ is a fixed, although unknown, value) can only be 0 or 1. Thus we can only wonder whether p may be assumed as a reasonable subjective probability, objectively based. That a *subjective* probability can be assigned to the statement $a(\boldsymbol{x}) \leq \theta \leq b(\boldsymbol{x})$ – an uncertain event – is doubtless (cf. for example Lindley, 1985, chap. 2). Jaynes (1976, p. 209) expresses the concept very clearly: "Indeed, isn't a matter of the most elementary common sense to recognize that, in the specific problem at hand, θ is just an unknown constant?". We'll see that the confidence coefficient p is just the better choice.

Most Bayesian textbooks deal with this topic; as an example, we shall quote the following passage from Bernardo and Smith (1994), as it expresses the position with great clarity: "... if we define a statistical procedure to consist of producing the interval $x \pm 1.96/\sqrt{n}$ whenever a random sample of size n from $N(x|\mu,1)$ is obtained, we are producing an interval which will include the true value of the parameter 95% of the time, *in the long run*. Note that this says *nothing* about the probability that μ belongs to the interval for any *given* sample" (p. 453). Further on, we read of an analogous interpretation of the upper confidence limit for a parameter θ with confidence coefficient $1 - a$, given $\overline{\theta}^{\alpha}(\boldsymbol{x})$: "Whether or not the *particular* $\overline{\theta}^{\alpha}(\boldsymbol{x})$ which corresponds to the observed data $\boldsymbol{x}$ is smaller or greater than θ is *entirely uncertain*" (p. 466).

Now, let us fix the attention to the first interval $I(\boldsymbol{X})$, just quoted, with coefficient 95%. Before the sample is drawn, we can say that our subjective probability – coinciding with the objective probability – that the interval obtained covers θ is 0.95; after the interval is read – for example, as [0.98,1.35], *no supplementary information* on the ability of this interval of covering θ *has been gained*; thus our degree of belief that $\theta \in I(\boldsymbol{X})$ cannot be changed after the sample has been observed. In other words, we attach the subjective probability 0.95 to the belief that an event having objective probability 0.95 has occurred.

It must be stressed that one thing is to tell a probability for a sentence like «$0.98 \leq \theta \leq 1.35$» (a purely subjective probability is admitted), and a totally different thing is to tell a probability for the sentence «$0.98 \leq \theta \leq 1.35$, being such interval obtained by a random experiment which yields intervals covering θ with probability 0.95»; in this case, the probability is still – and inevitably – of the subjective kind, but it can hardly be at variance with the objective probability related to the random interval $I(\boldsymbol{X})$ (cf. Frosini, 1989 and 1996*b*). Such viewpoint is perfectly coherent with the usual elicitation of subjective probabilities, suggested by the Bayesian School, in terms of odds of a wager: if I am willing to bet 95 dollars against 5 that a random interval includes the unknown θ, I do not see any reason why to change the terms of the bet once the random experiment has been effected and the interval computed, as the knowledge of the experimental results does not minimally alter the information available before the experiment was made.

6. SOME COUNTER-COUNTER-EXAMPLES

Example 1 – (Berger-Wolpert, 1988, p. 5): "Suppose X_1 and X_2 are independent and

$$P_\theta(X_i = \theta - 1) = P_\theta(X_i = \theta + 1) = 1/2 \qquad i = 1,2$$

$[-\infty < \theta < \infty]$ ... A 75% confidence set of smallest size for θ is

$C(X_1,X_2)$ = the point $(X_1 + X_2)/2$ if $X_1 \neq X_2$
= the point $X_1 - 1$ if $X_1 = X_2$

Notice, however, that when $x_1 \neq x_2$ it is *absolutely certain* that $\theta = (x_1 + x_2)/2$ ... Thus, from a post-experimental viewpoint, one would say that $C(x_1,x_2)$ contains θ with "confidence" 100% when $x_1 \neq x_2$, but only with "confidence" 50% when $x_1 = x_2$ ... Does it make sense to report a pre-experimental measure when it is known to be misleading after seeing the data?"

Here, what is misleading is the use of a correct model probability for inferential purposes, in the presence of a piece of information which deductively conditions the inference. In fact, it is known *before* the performance of the experiment – and *not after seeing the data* - that, if $X_1 \neq X_2$, we get $\theta = (X_1 + X_2)/2$ *by means of deduction*; as a consequence, the above "confidence set" cannot be used in inference, as it ignores – by gross negligence – an important piece of information. The correct way of expressing an inference (from a frequentist viewpoint) can only be in the following terms: (a) if $X_1 \neq X_2$, then $\theta = (X_1 + X_2)/2$ by deduction; (b) if $X_1 = X_2$, then $\{\theta = X_1 - 1\}$ and $\{\theta = X_1 + 1\}$ are both (conditional) confidence sets, each one including only one point, with confidence coefficient 0.50 (comment resumed from Frosini, 1993*a*, p. 371).

Example 2 – (Berger, 1980, p. 19). If X_1, ..., X_n are i.i.d. with uniform density in $(\theta - 1/2,\ \theta + 1/2)$, calling $T = (X_{(1)} + X_{(n)})/2$, with $X_{(1)} = \min\{X_i\}$ and $X_{(n)} = \max\{X_i\}$, a confidence interval with confidence coefficient $1\text{-}\alpha$ is

$$I(\boldsymbol{X}) : \left(T + \frac{\sqrt[n]{\alpha}}{2} - \frac{1}{2}, T - \frac{\sqrt[n]{\alpha}}{2} + \frac{1}{2} \right) \tag{8}$$

If $\alpha = 0.05$, $n = 25$, $x_{(1)} = 3$ and $x_{(25)} = 3.96$, the 95% confidence interval gives: $3.424 < \theta < 3.536$. However, it is certain that θ is included in the following interval:

$$\Theta^S : (X_{(n)} - 1/2 \leq \theta \leq X_{(1)} + 1/2); \tag{9}$$

in the case at hand this means that $3.46 \leq \theta \leq 3.50$, i.e. θ certainly belongs to a subinterval of the above 95% confidence interval. Berger (1980, p. 19) says that the conclusion of the classical procedure "seems ridiculous, in light of our certain knowledge that θ is in the smaller interval".

Berger is perfectly right in his judgment of the above statement; but such statement does not belong to "classical statistics", and is not ascribable to any reasonable person. Like the previous example, a model probability is used without making allowance for an important piece of information, which deductively changes the probability interval (8). In fact, the inequality (9), obtained *by deduction*, is known *before* the performance of the experiment, and any sensible inference procedure must take it into account. One way of doing this is to consider the intersection of (8) and (9) (cf. formula (6 bis)), namely the interval

$$\max\left(T+\frac{\sqrt[n]{\alpha}}{2}-\frac{1}{2}, X_{(n)}-\frac{1}{2}\right) \le \theta \le \min\left(T-\frac{\sqrt[n]{\alpha}}{2}+\frac{1}{2}, X_{(1)}+\frac{1}{2}\right) \qquad (10)$$

which is again a $(1 - \alpha)$ probability interval, now a real confidence interval suitable for inferential purposes. This confidence interval is *correct*, in that the information yielded cannot be improved by means of deduction; on the contrary, (8) is a probability interval having an abstract validity as a mathematical probability in the model, but *incorrect* if used for inferential purposes (comment resumed from Frosini, 1993*a*, p. 372).

For the case $n = 2$ this example was introduced by Welch (1939) in a famous paper, which "has been a warning for generations of statisticians *against* the use of conditional inference, in that it shows that a reasonable conditional inference, based on a highly informative ancillary statistic (the sample range $X_{(2)} - X_{(1)}$), is beaten by an apparently less informative procedure, which completely disregards the information provided by the ancillary statistic" (Frosini, 1993*b*, pp. 42, 44).

The above exposition has been a bit lengthy, owing to the remembrance of a knowledgeable American statistician who kindly informed me that interval (10) "is not reasonable" as a confidence interval.

Example 3 – (Jaynes, 1976, pp. 196-200). Jaynes exposes an example, similar to the above two in that it uses a support of the random variable depending on θ, however much more realistic. The problem is to estimate a location parameter θ, from the sample values $\{x_1, \ldots, x_n\}$ distributed according to the density

$$\begin{aligned} f(x \mid \theta) &= \exp(\theta - x) \quad && \text{for } x > \theta \\ &= 0 && \text{for } x < 0. \end{aligned}$$

First of all, Jaynes considers a (putative) confidence interval based on the distribution of the sample mean $\bar{X}$ (as $\bar{X}$ - 1 is an unbiased estimator of θ), and shows, for a particular case of three observations, that the numerical 90% interval for θ, thus obtained, "*lies entirely in the region* $\theta > x_1$, [with x_1 = the least value observed] *where it is obviously impossible for* θ *to be!*". Quite to the contrary, a reasonable Bayesian solution takes expressly into account the above restriction on the support, and is a function of the least value x_1, so that the posterior density for θ is positive for $\theta < x_1$, and equal to zero for $\theta > x_1$. The author, while assuming the above probability interval as a valid confidence interval for inference purposes (a gratuitous assumption), admits that it is a poor confidence interval, and acknowledges that a better choice, *leading to the same Bayesian interval*, is the confidence interval based on the least observation x_1, which is a sufficient statistic for θ (p. 199).

Jaynes, developing some general considerations from the above case, observes that "whenever the confidence interval is not based on a sufficient statistic, it is possible to find a 'bad' subclass of samples, *recognizable from the sample*, in which use of the confidence interval would lead us to an incorrect statement more frequently than is indicated by the confidence level; and also a recognizable 'good'

subclass in which the confidence interval is wider than it needs to be for the stated confidence level". Owing to the inductive and optimal properties shared by sufficient statistics, this warning by Jaynes, towards a preferred use of sufficient statistics, is worthy of complete approval. Anyway, his viewpoint is too optimistic; recourse to sufficient statistics does not exempt us from checking possible deductive implications of the (5) kind, as the next example will show.

Example 4 – This "counter-example", with ensuing "counter" comment, is not taken from other authors, and it seems realistic at least as the preceding Jaynes' example; formally, it could be included in the general topic of confidence sets with restricted parameter space (Cox-Hinkley, 1974, pp. 224-228). Its rationale comes from the same field of quality control of industrial devices, where the relevant random variable is of the normal type $N(\mu,\sigma^2)$, however with parameters subject to physical constraints; for simplicity, we'll comment only on the simple – though realistic – case of $\mu_0 \leq \mu \leq \mu_1$ and σ^2 known (of course, μ_0, μ_1 and σ^2 are practically fixed to some degree of approximation, they can never be considered as *exact* real numbers). Now, if a random sample of size n is obtained from $X \sim N(\mu,\sigma^2)$, the realization of the usual probability interval based on the *sufficient statistic* $\overline{X} \sim N(\mu,\sigma^2/n)$ is $[\overline{x} - z_p\, \sigma/\sqrt{n}\,,\ \overline{x} + z_p\, \sigma/\sqrt{n}\,] = [\overline{x} - c_p,\ \overline{x} + c_p]$, where z_p is the p-th percentile of the standard normal N(0,1), if we want a coverage probability of $(2p - 1)$. Now, for any $\mu_0 \leq \mu \leq \mu_1$, the probability statement

$$P(-c_p \leq \overline{X} - \mu \leq c_p) = 2p - 1$$

is perfectly valid, and is also valid – as deductively equivalent – the probability statement

$$P(\overline{X} - c_p \leq \mu \leq \overline{X} + c_p) = 2p - 1.$$

If we pass from a *probabilistic* statement to an *inductive* statement, all the certain facts to our knowledge must be taken into account, thus a confidence (subjective probability) of $2p - 1$ is attached to the interval of μ values

$$\max(\overline{x} - c_p\,,\ \mu_0) \leq \mu \leq \min(\overline{x} + c_p,\ \mu_1)$$

showing that, not taking account of prior knowledge about the parameter space, we could determine an interval which includes (practically) impossible μ values.

Example 5 – (Howson and Urbach, 1993, p. 208). This example envisages a consignment of tulip-bulbs; for simplicity, only two hypotheses are considered: 40% red-flowering as the null hypothesis h_1, 60% red-flowering as the alternative hypothesis h_2. The following is an abridged table from Table VIII in Howson and Urbach (p. 208); the "minimum proportion" is the minimum proportion of red tulips in the sample, needed to reach h_1 at the 5% level.

Sample size n	minimum proportion	power against h_2
10	0.70	0.37
20	0.60	0.50
50	0.50	0.93
100	0.48	0.99
1000	0.426	1.0

Howson and Urbach write: "It will be noticed that as n increases, the critical proportion of red tulips that would reject h_1 at the 0.05 level approaches more closely to 40 per cent, that is, to the proportion which h_1 asserts is contained in the consignment. Bearing in mind that the only alternative to h_1 that is admitted in this simple example is that the consignment contains red tulips in the proportion of 60 per cent, an unprejudiced consideration of these data would, it seems to us, lead to the conclusion that as n increases, these so-called critical values *support* h_1 more and more".

Leaving aside this last affirmation (Neyman-Pearson tests are not conceived to give support to any hypothesis, but to help choose between the hypotheses, given a comparison of the consequences of the possible errors), this analysis is formally correct, but it assumes the existence of a Non-Bayesian statistician who is so insane as to maintain a standard value $a = 0.05$ irrespective of the sample size and the power of the test. It is well known (since the first works by Neyman and Pearson) that the respective values of the error probabilities α and β must be carefully determined according to the possible consequences of the errors; the ratio α/β is usually fixed in advance, and possibly fulfilled (approximately), compatibly with a sample size not too small. As far as two *scientific* hypotheses are concerned (not the bulb case), the best choice seems generally to put both hypotheses on the same footing, thus $\alpha = \beta$ (or approximately so), irrespective of n.

Example 6 – Jeffreys-Lindley paradox (Jeffreys, 1939, 1948; Berger, 1985, pp. 150-151).

This example, so much discussed in the literature, stretches the same problem already commented on in the preceding example to the limit. It must be acknowledged that many cook-books – following the Neyman-Pearson approach – do not carefully concern the choice and ratio between the error probabilities α and β, or – generally speaking – the power function. It is quite possible that some coarse practitioners (not statisticians) do not realize that the consistency property of inference procedures implies – when the null hypothesis is a *point* hypothesis and the alternative is a very comprehensive *composite* hypothesis – that standard levels of the error probability α tend to almost always accepting the null hypothesis for small n (sample size), and rejecting it for large n.

Most null hypotheses envisaged in elementary textbooks are point hypotheses; for example: (a) the mean of a continuous random variable is equal to 5 (or another *exact* real number, e.g. π), (b) two or more random variables are independent, (c) the distribution of a certain characteristic is normal, or Poisson etc. (cf. Frosini, 2001, p. 374; 2004, p. 276). The common feature of these hypotheses is that they determine a dimensionless *point* in a space (of reals, of vectors, of func-

tions etc.); they are *model hypotheses*, and are *necessarily false* (at least at a human level) if referred to real phenomena. Thus, the above implication of consistency is totally correct: when the collected information increases, recognition of the falsity of the null hypothesis becomes easier and easier. All this means that classical tests of (point null) hypotheses are *practically* acceptable only for "small" sample sizes, where *small* is to be deemed according to the precision of the random variables involved. When the sample size is small, the sampling variability of the test statistic is generally so large as to dominate the imprecise (being *too precise*) specification of the null hypothesis. As we let the sample size increase, we must acknowledge the growing unsuitableness of the test procedure in order to answer the practical problem in the real world. Among the possible solutions: (a) avoid applications of such tests in case of large samples, and limit the inference to estimation procedures; (b) restate the problem in more acceptable (practical) terms, e.g. by fixing intervals for parameters, for example: $H_0 = \theta \in [a,b]$.

Now, let us come to the Jeffreys-Lindley paradox. A generally correct Bayesian way of dealing with testing two incompatible hypotheses H_0 and H_1, is to assign respective prior probabilities π_0 and π_1, possibly spread out over the simple hypotheses constituting H_0 and H_1. This is *mathematically* possible for every conceivable way of specifying the two hypotheses, but loses any justification for the case here considered, i.e. H_0 being a point hypothesis and H_1 the complement with respect to the whole space of admissible hypotheses, for the simple fact that the only reasonable assignments are $P(H_0) = 0$ and $P(H_1) = 1$. The specific example worked out by Bayesians (see e.g. Berger, 1985, pp. 150-151) refers to the usual normal distribution with known variance, mean $\mu = \mu_0$ as the null hypothesis, and $\mu \neq \mu_0$ for the alternative, with prior such that $\pi_0 = 1/2$ is assigned to the point μ_0, and $\pi_1 = 1/2$ is spread out over the whole real line, excepting μ_0.

Let it be sufficient to comment on the following calculations (Berger, 1985, p. 151; Frosini, 2004, p. 284): by fixing $z = 1.96$ for the standardized value of the sample mean ($P = 0.05$) of a sample size n, Berger obtains posterior probabilities for H_0 increasing from 0.35 when $n = 1$ to 0.80 when $n = 1000$; hence he observes that "classical theory would allow one to reject H_0 at level $\alpha = 0.05$ ... But the posterior probability of H_0 is quite substantial, ranging from about 1/3 for small n to nearly 1 for large n. Thus $z = 1.96$ actually provides little or no evidence against H_0 (for the specific prior)". In this way, a comparison is made between an absurd application of the Neyman-Pearson approach (α fixed for every n) and an absurd application of the Bayesian approach (a finite and very substantial probability for a point hypothesis): whatever the result of this comparison, I cannot see any interest in it. Quite to the contrary, if one starts with a small but non-degenerate interval $[-b,b]$ for θ in H_0, everything goes well. Although allowing different conclusions according to different interval widths and prior distributions (which is correct), no «astonishing comparisons» can be obtained, and there are reasonable choices of width and prior which ensure sensible agreement between Neyman-Pearson and Bayesian approaches (Frosini, 2004, pp. 282-284). For example, by the choice of $\pi_0 = 0.1$ and $b = 0.1$, one obtains the following values for the posterior probabilities:

n	1	10	20	50	100	1000
Post.	0.054	0.052	0.058	0.064	0.065	0.064

In this case the posterior probability of H_0 is around 0.06 from $n = 1$ to $n = 1000$, thus ensuring an exceptional agreement with respect to the nominal value of the probability $\alpha = 0.05$, established by the Neyman-Pearson approach (perhaps you will not believe, but it is the only case with $\pi_0 \neq 1/2$ that I have tried).

Supplementary Remarks

Two Bayesian scholars, named for convenience *A* and *B*, have kindly accepted to read the paper and make some observations, hopefully for gaining formal or substantial improvements; in any case, I must gratefully thank them for their willingness and sincere (even crude) examination. Perhaps they are not representative of the wide population of Bayesians (although I surmise they are); anyway, I must peacefully take note of their disagreement about the approach and the whole contents of the paper. A few supplementary remarks may be suitable in a final tentative of making me understand.

Unfortunately, Bayesian *A* did not arise any specific point in support of a general and complete disagreement, however clearly expressed; thus no counter-remarks can be made. Instead, Bayesian *B* developed some discussion on a few points, thus allowing to understand where is the real or presumed reason of disagreement.

As a first point, Bayesian *B* clearly states his creed about the Bayesian approach to inference: "Bayesians consider the parameter just a device to better deal with the representation theorem. The real thing is the posterior predictive distribution of the "next observation". Of course there are situations where the parameter is fixed; actually, the very meaning of a model relies on that. However the prior reflects the information of the researcher about the possible values of the parameter and this is done via the probability language". Perfectly said.

About the comments on finite population sampling, contained in Section 3 of the paper, Bayesian B correctly observes that "there is nothing Bayesian in Godambe argument"; however, he continues saying that "the real issue is the conflict between the repeated sampling principle and the conditionality principle". Two counter-remarks: (1) Although a reasonable principle in itself, it is not a postulate accepted by non-Bayesian statisticians, thus the observation has a limited value only inside the Bayesians. (2) Most of all, the above observation has nothing to do with my comments, whose aim is clearly – and exclusively – to counter Godambe's assertions, namely showing (a) the existence of an ordinary and non-flat likelihood, and (b) the absurdity of an "estimator" which - to be applied - requires the knowledge of the whole population.

Concerning the basic tenet in Section 4, i.e. that "a confidence interval aimed at making inference on the parameter" must take account of "all certain facts and

constraints" known to the research worker (from Summary), Bayesian *B* observes: "But this is exactly what the likelihood function (not the Bayes theorem) does. In Example 2 (Berger, 1980, p. 19, but see also Cifarelli and Muliere, Statistica bayesiana, Iuculano, 1989) the constraints on the parameter are provided by the observed sample (that is, by the likelihood function)".Counter-remarks: (1) The exclusion of the Bayes theorem is rather curious: perhaps the prior distribution (beside the likelihood) does not necessarily include all the assumed knowledge? (2) Perhaps the paper states something to the contrary? The observation of Bayesian *B* has nothing to do with my paper; it seems that he has read another paper.

On this same point, Bayesian *B* continues: "It is true that one could, in principle, take into account all the possible sets of constraints induced by all the possible samples; but then the real coverage of the procedure would be much much much more difficult than the "*trivial equality*" 6 bis!". This conclusion is simply false, in three senses: (1) there are usually no constraints, or they are of a very limited number and kind. (2) There is no need to forecast "all the possible sets of constraints induced by all the possible samples". (3) If such a requirement would be judged really necessary, and the "possible sets of constraints" on the parameter should constitute a large set, these same constraints should be considered in establishing the likelihood function; thus, where is the criticism?

About the basic tenet of counter-example 6, Bayesian *B* makes the following interesting observations: "The role of point null hypothesis: I believe that Bayesian, frequentist and likelihood people, all agree in dealing with the simple estimation problems considered in this paper. Real concern may only arise in testing scenario if the statistician refuses the paramount role of point null hypothesis in testing. Even though we formalize problems with the aid of some Greek letter like θ, they are almost always physical quantities and their values have a specific meaning. In a famous experiment in the 1920's Eddington compared the Newtonian theory and the new (and not yet observed at that time!) Einsteinian relativity theory in an indirect way; he measured the distance of a specific body from the earth; the two theories would have been imply different distance from the earth and few measurements were enough to support the relativity theory. The presence of significant specific values of the parameter is perhaps one of the main reasons why (social and hard) science needs statistics". Excepting minor points of disagreement, which are not worth the effort, Bayesian *B* has to be congratulated for a clear and convincing statement about the usefulness of point null hypotheses. But his observations are wholly unrelated to the content of the paper. Perhaps the cited astronomical measure was conceived as a precise real number? Of course not, it would be ridiculous! Even, it was certainly not expressed in centimetres, or inches. And the comparison problem was not conceived as an *exact* Bayesian test, putting a prior finite probability on a fixed measure (which is precisely the kind of problem examined in counter-example 6).

Dipartimento di Scienze Statistiche BENITO VITTORIO FROSINI
Università Cattolica del Sacro Cuore di Milano

REFERENCES

M.J. BAJARRI, , AND J.O. BERGER, (2004), *The Interplay of Bayesian and Frequentist Analysis*, "Statistical Science", 19, pp. 58-80.

J.O. BERGER, (1980), *Statistical Decision Theory*, Springer, New York.

J.O. BERGER, (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer, New York.

J.O. BERGER, (2006), *The Case for Objective Bayesian Analysis*, "Bayesian Analysis", 1, pp. 385-402.

J.O. BERGER, AND R.L. WOLPERT (1988), *The Likelihood Principle* (Second Edition), Institute of Mathematical Statistics, Hayward.

J.M. BERNARDO, (1979), *Reference Posterior Distributions for Bayesian Inference*, "Journal of the Royal Statistical Society B", 41, pp. 113-147.

J.M. BERNARDO, (2005), *Reference Analysis*, "Handbook of Statistics", Vol. 25, pp. 17-90, Elsevier, Amsterdam.

J.M. BERNARDO, AND A.F.M. SMITH, (1994), *Bayesian Theory*, Wiley, Chichester.

R. CARNAP, (1962), *Logical Foundations of Probability* (Second Edition), The University of Chicago Press, Chicago.

D.R. COX, (1958), *Some Problems Connected with Statistical Inference*, "The Annals of Mathematical Statistics", 29, pp. 357-372.

D.R. COX, (1986), *Some General Aspects of the Theory of Statistics* , "International Statistical Review", 54, pp. 117-126.

D.R. COX, D.V. HINKLEY, (1974), *Theoretical Statistics*, Chapman and Hall, London.

B. EFRON, (1986), *Why Isn't Everyone a Bayesian?*, "The American Statistician", 40, pp. 1-5.

B. EFRON, AND C. MORRIS, (1971), *Limiting the Risk of Bayes and Empirical Bayes Estimators – Part I: The Bayes Case*, "Journal of the American Statistical Association", 66, pp. 807-815.

R.A. FISHER, (1934), *Two New Properties of Mathematical Likelihood*, "Proceedings of the Royal Society", *A*, 144, 285-307.

R.A. FISHER, (1935), *The Logic of Inductive Inference (with discussion)*, "Journal of the Royal Statistical Society", 98, Pt. 1, pp. 39-82.

B.V. FROSINI, (1989), *La Statistica metodologica nei convegni della SIS*, Società Italiana di Statistica, *Atti del Convegno "Statistica e Società*", pp. 197-228, Pacini. Pisa.

B.V. FROSINI, (1991), *On Some Applications of the Conditionality Principle*, "Statistica Applicata", 3, pp. 555-568.

B.V. FROSINI, (1992), *Sui racconti con la morale, ovvero come rifiutare l'uso di informazioni certe*, "Statistica Applicata", 4, pp. 33-43.

B.V. FROSINI, (1993*a*), *Likelihood Versus Probability*, International Statistical Institute, *Proceedings of the ISI 49-th Session*, Vol. 1, pp. 359-376, Firenze.

B.V. FROSINI, (1993*b*), *Global and conditional tests*, "Metron", 51, pp. 27-58.

B.V. FROSINI, (1996*a*), *Likelihood, Superpopulation and Other Problems in Sampling from Finite Populations*, Società Italiana di Statistica, *100 Anni di Indagini Campionarie*, pp. 217-239, CISU, Roma.

B.V. FROSINI, (1996*b*), *Some Reasons for Reconciling Confidence Intervals and Bayesian Intervals*, "Statistica", 56, pp. 301-311.

B.V. FROSINI, (1999), *Conditioning, Information and Frequentist Properties*, "Statistica Applicata", 11, pp. 165-184.

B.V. FROSINI, (2001), *Metodi Statistici*, Carocci, Roma.

B.V. FROSINI, (2004), *On Neyman-Pearson Theory: Information Content of an Experiment and a Fancy Paradox*, "Statistica", 64, pp. 271-286.

B.V. FROSINI, (2005), *Objective Bayesian Intervals: Some Remarks on Gini's Approach*, "Metron", 63, pp. 435-450.

V.P. GODAMBE, (1955), *A Unified Theory of Sampling from Finite Populations*, "Journal of the Royal Statistical Society B", 17, pp. 269-278.

V.P. GODAMBE, (1965), *A Review of the Contributions Towards a Unified Theory of Sampling from Finite Populations*, "Review of the International Statistical Institute", 33, pp. 242-258.

V.P. GODAMBE, (1969), *A Fiducial Argument with Applications to Survey Sampling*, "Journal of the Royal Statistical Society B", 31, pp. 246-260.

C. HOWSON, AND P. URBACH, (1993), *Scientific Reasoning: The Bayesian Approach* (Second Edition), Open Court, Chicago.

E.T. JAYNES, (1968), *Prior Probabilities*, "IEEE Transactions on System, Science and Cybernetics", SCC-4 (September 1968), pp. 227-241.

E.T. JAYNES, (1976), *Confidence Intervals vs Bayesian Intervals*, Harper and Hooker (eds), "Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science", Vol. II, pp. 175-257.

H. JEFFREYS, (1939/1948), *Theory of Probability*, Oxford University Press, Oxford.

B.W. LINDGREN, (1962), *Statistical Theory*, Macmillan, New York.

D.V. LINDLEY, (1985), *Making Decisions* (Second Edition), Wiley, London.

M. MANDELKERN, (2002), *Setting Confidence Intervals for Bounded Parameters (with discussion)*, "Statistical Science", 17, pp. 149-172.

R. ROYALL, (1968), *An Old Approach to Finite Population Sampling Theory*, "Journal of the American Statistical Association", 63, pp. 1269-1279.

B.L. WELCH, (1939), *On confidence limits and sufficiency, with particular reference to parameters of location*, "The Annals of Mathematical Statistics", 10, pp. 58-69.

SUMMARY

On some counter-counter-examples about classical inference

This paper deals with theoretical concepts and practical examples, aimed at showing that non-Bayesian inference is liable to result in mistakes or unacceptable conclusions, and proves that they are not justified. Section 2 comments on examples when an objective prior distribution exists, and shows how widely one can be mistaken in using a prior quite distant from the real one. Section 3 comments on two results by Godambe, stressing that – in sampling from finite populations – no flat likelihood exists, while an unbiased linear "estimator" with zero variance does not exist, unless we reach a complete knowledge of the population. Section 4 stresses the fundamental difference between a "probability interval" for a parameter, and a "confidence interval" aimed at making inference on the parameter, thus summarizing all certain facts and constraints able to shrink such an inferential interval. Section 5 explains why we are justified in attaching an inductive meaning to a realized confidence interval. Finally, Section 6 counters some well known counter-examples spread in the Bayesian literature, showing that they are unacceptable from a sound inductive basis.